**##################################################################################**

**########## BI-GRAMS AND TRI-GRAMS FOR SPEECHES OF 39TH CONGRESS ###################**

**################### Reading the data from file PrelimData.RData #########################**

**##################################################################################**

**rm(list = ls())**

**library(tm)**

**library(slam)**

**library(wordcloud)**

**## LOADING DATA (39th Congress)**

**load("C:\\Johannes Ledolter\\2020March01Book\\Chapter10WEB\\PrelimData.RData")**

**################################################################################**

**## CREATING THE CORPUS AND THE RESULTS IN CHAPTER 10 #############################**

**################################################################################**

**corpus <- VCorpus(VectorSource(data),readerControl = list(reader = readPlain))**

**corpus1 <- tm\_map(corpus, stripWhitespace)**

**corpus2 <- tm\_map(corpus1, content\_transformer(tolower))**

**corpus3 <- tm\_map(corpus2, removePunctuation)**

**corpus4 <- tm\_map(corpus3, removeNumbers)**

**corpus5 <- tm\_map(corpus4, removeWords, stopwords("english"))**

**## here we are looking at all 97,000 plus speeches (short or long)**

**corp.dtm <- DocumentTermMatrix(corpus5,control=list(stemming=FALSE)) ## no stemming is the default**

**dim(corp.dtm)**

**findFreqTerms(corp.dtm,500)**

**## Note: words "hrno" and "sno" reference bill numbers, such as "(H.R.No.30)" and "(S.No.56)"**

**## as numbers and special characters such as . ( ) are stripped from text**

**## looking at frequencies of combined (coded) phrases**

**sum(as.matrix(corp.dtm[,"rfg"]))**

**sum(as.matrix(corp.dtm[,"declarindep"]))**

**sum(as.matrix(corp.dtm[,"postoffice"]))**

**sum(as.matrix(corp.dtm[,"civilwar"]))**

**sum(as.matrix(corp.dtm[,"newyork"]))**

**sum(as.matrix(corp.dtm[,"rhodeisland"]))**

**sum(as.matrix(corp.dtm[,"abife"]))**

**sum(as.matrix(corp.dtm[,"gentfrom"]))**

**sum(as.matrix(corp.dtm[,"senfrom"]))**

**sum(as.matrix(corp.dtm[,"friendfrom"]))**

**sum(as.matrix(corp.dtm[,"unitedstates"]))**

**## BI-GRAMS**

**BigramTokenizer <- function(x)**

**unlist(lapply(ngrams(words(x), 2), paste, collapse = " "), use.names = FALSE)**

**bi2.dtm <- DocumentTermMatrix(corpus5, control = list(tokenize = BigramTokenizer))**

**bi2.dtm**

**## displaying bigram frequencies**

**findFreqTerms(bi2.dtm,1000) ## occurring more than 1000 times**

**## Note: bigrams "bill hrno" and "bill sno" reference bill numbers as "bill (H.R.No.30)" and "bill (S.No.56)"**

**## as numbers and special characters such as . ( ) are stripped from text**

**## frequency distribution**

**ffff=rollup(t(as.DocumentTermMatrix(bi2.dtm)),2,na.rm=TRUE,FUN=sum)**

**freq=as.vector(ffff)**

**labels=rownames(ffff)**

**wf=data.frame(labels,freq)**

**wf=wf[order(-wf$freq),]**

**wf[1:10,]**

**library(ggplot2)**

**p=ggplot(subset(wf,freq>2000),aes(labels,freq))**

**p=p+geom\_bar(stat="identity")**

**p=p+theme(axis.text.x=element\_text(angle=45,hjust=1))**

**p**

**## displaying word clouds**

**set.seed(142)**

**dark2 <- brewer.pal(6,"Dark2")**

**wordcloud(labels,freq,scale=c(2,.5),max.words=50,rot.per=0.1,colors=dark2)**

**## TRI-GRAMS**

**BigramTokenizer <- function(x)**

**unlist(lapply(ngrams(words(x), 3), paste, collapse = " "), use.names = FALSE)**

**bi3.dtm <- DocumentTermMatrix(corpus5, control = list(tokenize = BigramTokenizer))**

**bi3.dtm**

**## displaying bigram frequencies**

**findFreqTerms(bi3.dtm,100)**

**## frequency distribution**

**ffff=rollup(t(as.DocumentTermMatrix(bi3.dtm)),2,na.rm=TRUE,FUN=sum)**

**freq=as.vector(ffff)**

**labels=rownames(ffff)**

**wf=data.frame(labels,freq)**

**wf=wf[order(-wf$freq),]**

**wf[1:10,]**

**library(ggplot2)**

**p=ggplot(subset(wf,freq>2000),aes(labels,freq))**

**p=p+geom\_bar(stat="identity")**

**p=p+theme(axis.text.x=element\_text(angle=45,hjust=1))**

**p**

**## displaying word clouds**

**set.seed(142)**

**dark2 <- brewer.pal(6,"Dark2")**

**wordcloud(labels,freq,scale=c(2,.5),max.words=50,rot.per=0.1,colors=dark2)**

**## TARGETED ANALYSIS**

**## transferring document term matrices of counts into document matrices of occurrences**

**Bcorp.dtm=weightBin(corp.dtm)**

**Bbi2.dtm=weightBin(bi2.dtm)**

**Bbi3.dtm=weightBin(bi3.dtm)**

**## meta variables**

**meta2**

**year=as.numeric(meta5)**

**yearcat=factor(year)**

**table(yearcat)**

**perc=dim(4)**

**## negro**

**indterm=as.matrix(Bcorp.dtm[,"negro"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**names(perc)[1]="1865"**

**names(perc)[2]="1866"**

**names(perc)[3]="1867"**

**names(perc)[4]="1865+1866"**

**perc**

**## slave**

**indterm=as.matrix(Bcorp.dtm[,"slave"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## indian**

**indterm=as.matrix(Bcorp.dtm[,"indian"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## war**

**indterm=as.matrix(Bcorp.dtm[,"war"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## sweat of his brow**

**indterm=as.matrix(Bbi2.dtm[,"sweat brow"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## indian war**

**indterm=as.matrix(Bbi2.dtm[,"indian war"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## black slave**

**indterm=as.matrix(Bbi2.dtm[,"black slave"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## negro slave**

**indterm=as.matrix(Bbi2.dtm[,"negro slave"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## indian slave**

**indterm=as.matrix(Bbi2.dtm[,"indian slave"])**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## post office (here we combine frequencies of bi-gram and the single word postoffice)**

**indterm1=as.matrix(Bbi2.dtm[,"post office"])**

**table(indterm1)**

**## postoffice**

**indterm2=as.matrix(Bcorp.dtm[,"postoffice"])**

**table(indterm2)**

**indterm=indterm1+indterm2**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**

**## republican form government (here we combine frequencies of tri-gram and the single word rfg)**

**indterm1=as.matrix(Bbi3.dtm[,"republican form government"])**

**table(indterm1)**

**## rfg**

**indterm2=as.matrix(Bcorp.dtm[,"rfg"])**

**table(indterm2)**

**indterm=indterm1+indterm2**

**indterm=indterm>0**

**table(indterm)**

**meta2[indterm]**

**sort(table(meta2[indterm]))**

**tt=table(indterm,yearcat)**

**tt**

**for (i in 1:3) {**

**perc[i]=tt[2,i]/(tt[1,i]+tt[2,i])**

**}**

**perc[4]=(tt[2,1]+tt[2,2])/(tt[1,1]+tt[1,2]+tt[2,1]+tt[2,2])**

**perc=round(perc,3)**

**perc**